Path Crawler：基于Python的路径抓取工具

Path Crawler简介

基于Python3.6和requests工具包，使用Web Map API获取路径数据。

安装

安装Python3.6

首先从https://www.python.org/ 选择对应系统的Python3.6下载并安装。（Windows用户直接点击这里下载 https://www.python.org/ftp/python/3.6.3/python-3.6.3.exe ）。推荐在Windows 10系统下安装和运行Python和该程序。

安装过程中，**注意确认勾选Add Python 3.6 to PATH**（如图）。安装成功即可。

获取Path Crawler程序

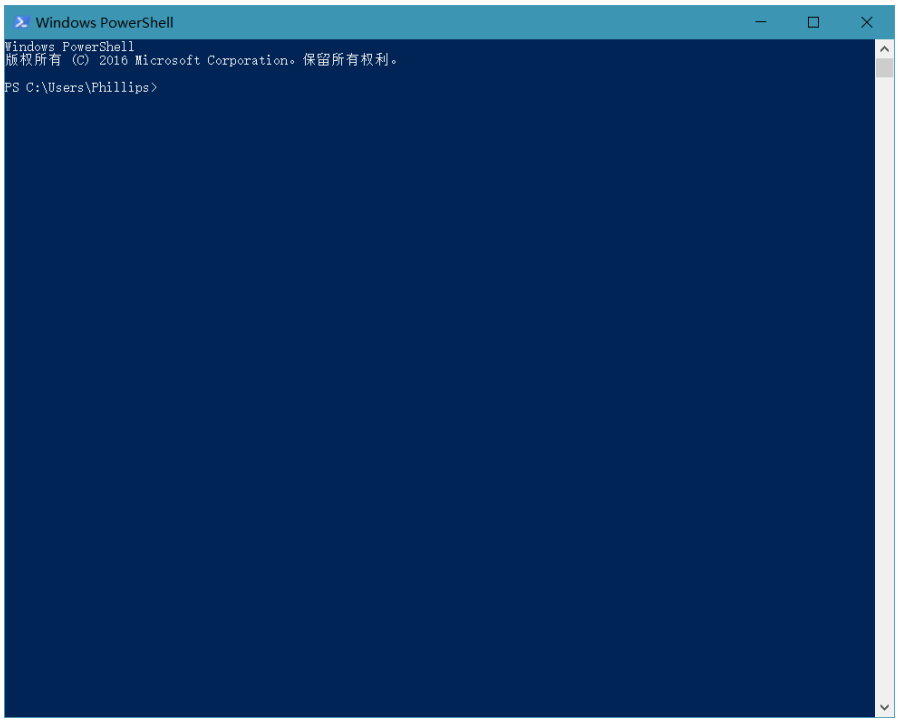
如果你会使用git，可以在想要放置程序的目录下，通过命令行输入以下代码：

```
git clone
https://github.com/esdream/PathCrawlerInPython.git
```

如果你不会使用git，可以通过这里Path Crawler下载压缩包，解压即可。

如何使用命令行工具？

如果你使用的是Windows，请在“开始”中搜索**PowerShell**，打开后如下图所示：



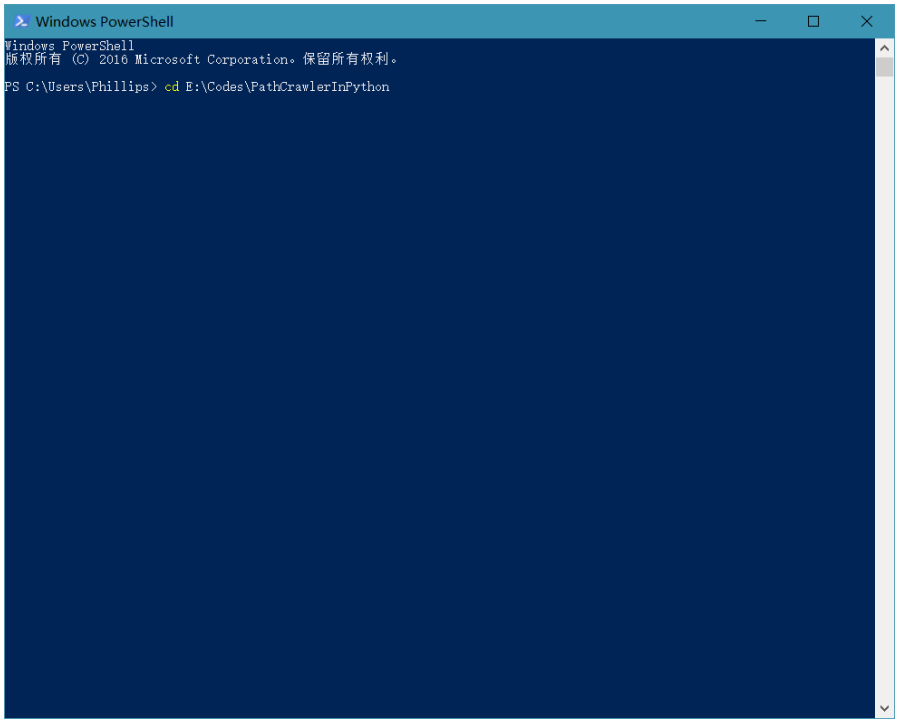
如果使用的是Windows系统，之后的各项涉及到命令行的操作均推荐使用PowerShell作为命令行工具，这在程序运行时效率更高。

使用命令行工具进入指定目录的方法

将目录的位置从地址栏处复制下来，打开命令行工具后，在命令行中输入

```
cd 目录
```

例如在Windows上使用PowerShell，且PathCrawlerInPython目录的路径为 `E:\Codes\PathCrawlerInPython` 。将目录路径复制后，在PowerShell中输入 `cd` ，然后输入一个空格，然后**点击右键**将 `E:\Codes\PathCrawlerInPython` 粘贴到命令行中，最后回车即可进入到该目录下。如图所示。



安装依赖包

获得Path Crawler程序后，打开命令行工具，进入PathCrawlerInPython目录。输入以下命令：

```
pip install -r requirements.txt
```

即可完成依赖包的安装。

如果以上命令没有正确安装依赖包，你可以在命令行中输入以下命令：

```
pip install requests
```

格式化OD数据

你可以从以下链接中下载不同交通方式OD文件的格式。这是一个Excel文件，每个sheet中是**sheet名对应的交通方式的OD文件格式**。

不同交通方式OD文件格式下载

你需要将你的Origin-Destination数据格式化为**按逗号隔开的`.csv`格式文件**。在Excel中，你可以通过“导出”实现这一功能。注意，**OD文件导出时一定要包含表头！**

你也可以使用其他工具或文本编辑器创建OD文件，格式要求同上。

抓取路径数据

格式化完成后，将OD文件放置到 `path_crawler/data/od/` 目录中。然后打开命令行工具，进入PathCrawlerInPython目录，输入以下命令：

```
python -m path_crawler.path_spider
```

Path Crawler中提供了两种Web地图API——百度API和高德API，提供了4种交通方式——驾车（百度，高德），公交（百度），骑行（百度）和步行（百度）。你可以根据OD数据的格式和需要，按照命令行的文字提示，输入对应的参数。

在输入参数过程中如果有输入错误，按**ctrl + c**结束命令，重新输入以上抓取命令即可。

如果命令行显示 `The OD file is not existed!` 。说明没有将OD数据放置到 `path_crawler/data/od/` 中。

如果命令行显示 `The path data have been crawled.` 。说明**输出文件已经存在**，可能该OD文件已经被抓取过。可以修改输出文件名或者将已经存在的输出文件拷贝或删除，重新输入抓取命令即可。

抓取到的路径数据被存储在 `path_crawler/data/path_data/` 目录中。路径数据是 `.db` 格式文件，可以下载sqlitestudio打开、查看、查询和导出结果。

处理错误

处理抓取错误

在1.5.1版本中，Path Crawler会自动处理抓取错误的文件，无需手动处理。

处理解析错误

如果OD数据某一条数据在抓取时出现了解析错误，则会被记录在 `input-filename.csv` 文件中并放置在 `path_crawler/data/parse_error/` 目录下。

一般情况下，出现解析错误的原因是该类型Web地图API或该种交通方式下，无法正确获取路径。你需要将这一csv文件复制至 `path_crawler/data/od/` 目录中并使用另一种Web地图API或交通方式抓取。

其他工具

地理编码工具

转换地址名称与坐标。

将地址转换为坐标

首先，将你的地址数据按照以下格式存储为 `.csv` 文件（以**UTF-8编码**），存储时包含表头。

ID	ADDRESS	CITY

然后将该文件放置到 `path_crawler/data/geo_encoding/` 目录中。

最后打开命令行工具，进入PathCrawlerInPython目录，运行以下命令：

```
python -m path_crawler.geo_encodings
```

运行完成后，编码完的数据存储在 `path_crawler/data/geo_encoding/` 中。